D-019

WWW 画像検索システムにおける検索質問拡張に基づくフィードバック検索
Feedback Retrieval based on the Query Expansion on WWW Image Retrieval Systems

竹安 真紀夫† 獅々堀 正幹† 柘植 覚† 北 研二‡
Makio Takeyasu Masami Shishibori Satoru Tsuge Kenji Kita

1. はじめに

現在，WWW 上には膨大な量の画像データが存在する．この膨大な量の画像データの中から一度の検索でユーザが所望する画像を取得することは困難である．そこで従来より，ユーザからのフィードバック情報を利用した検索システムが提案されている．一般に，WWW 画像検索システムにおけるフィードバック検索は，フィードバック情報を用いて検索結果をソーティングするものが主流であり，検索質問自体にフィードバック情報を反映させていなかった．

本稿では，WWW 画像検索において，フィードバック情報を用いて検索質問を拡張し，拡張された検索質問に対してフィードバック検索を行う手法を提案する．本手法は，ユーザが選択した正解画像にリンクする HTML ページ中の言語情報より，検索質問の類義語を特定する．この類義語を用いて再検索し，ユーザが選択した正解画像と類義語で検索した画像を比較し，内容的に類似した画像をユーザに出力する．

2. 従来のフィードバックモデル

画像検索システムにおける従来のフィードバックモデルは，ユーザが選択した画像と検索結果の画像を比較し，類似している画像を上位に入れ換えるものであった．しかし，検索画像数が多くなるにつれ，ノイズとなる画像が増加し検索精度の限界があった．これに対して，予め人手で登録した(検索単語の)類義語での検索結果をフィードバック情報に則して順位付けし，より多くの正解画像を検索する手法も提案されている[1]．しかし，用いる類義語がフィードバック情報に必ずしも即した単語であるとは限らず，また，類義語特定の自動化も必要であった．そこで，本稿では，フィードバック情報から自動的に類義語を特定する手法を用いる．

文書検索の分野では，適合性フィードバックを用いて検索質問を拡張する手法が一般に用いられる．これは適合文書に含まれる単語の重みを大きくし，不適合文書に含まれる単語の重みを小さくするように検索質問の修正を行う手法である[2]．WWW 画像検索システムにおいても，ユーザが正解と判断した画像にリンクするページを適合文書，不正解とした画像にリンクするページを不適合文書として検索単語を拡張することも考えられる．しかし，Google 等の WWW 画像検索システムは，画像近傍の言語情報から検索結果を表示しているため，画像内容とページの内容が異なる場合も見受けられる．このため，不正解画像であるにも関わらず，ページの内容が正解画像ページの内容と類似していたために，類義語となるべき単語を見落としてしまう可能性が高い．そこで，本稿では正解画像にリンクするページ内の言語情報のみを用いて検索単語を拡張し，既存の WWW 画像検索システムにおいてフィードバック検索を行う手法を提案する．

†徳島大学大学院工学研究科
‡徳島大学高度情報化基盤センター

3. 検索質問拡張に基づくフィードバック検索

3.1 本手法の概要

図 1 に本稿で提案するフィードバック検索手法を示し，手順を説明する．なお，手順 2 で示す類義語候補の重み付け，および手順 3 で示す検索質問の拡張方法については 3.2 で詳しく述べる．

手順 1: 正解画像の選択
検索単語を WWW 画像検索システムに入力し，上位 n 件からユーザの希望する正解画像 $\mathrm{Image}_i(1 \le i \le n)$を選択する．

手順 2: ページ内容の解析
Image_iにリンクする HTML ページを形態素解析し，出現単語 $\mathrm{w}_j(j \ge 1)$と重み $\mathrm{Weight}(\mathrm{w}_j)$を集計する．

手順 3: 検索単語の拡張
類義語候補単語 w_jから多義性のある一般的な単語を除去し類義語を特定する．

手順 4: 類似画像の特定
手順 3 で特定した類義語を WWW 画像検索システムに入力し，上位 m 件の検索結果 $\mathrm{Image}_k(1 \le k \le m)$と正解画像 Image_iとの類似度を計算する．

図 1 では，手順 1 において，検索単語に「松井」を入力し，「松井秀喜が野球している画像」を正解画像としている．次に，手順 2 で正解画像のページ内に含まれる単語を取得し，手順 3 において，「ヤンキース」等の類義語を特定する．最後に，これらの類義語を WWW 画像検索システムに入力した検索結果と手順 1 で選択した正解画像の類似画像検索を行い，結果を出力する．

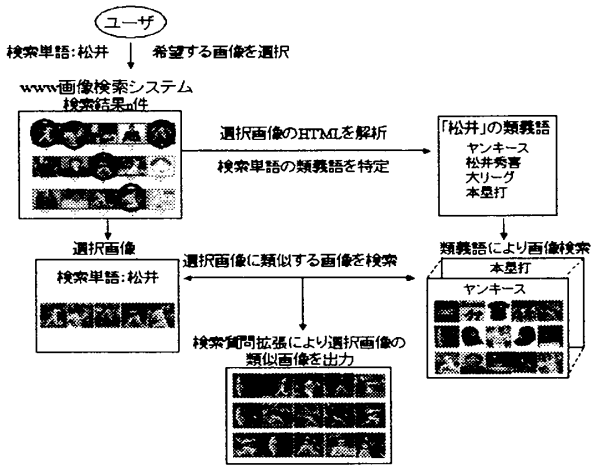

図 1： 本提案手法の概要

3.2 検索単語の類義語特定

検索単語の類義語を特定し，検索質問の拡張を行う．2. で述べたように画像近傍の言語情報により画像検索を行う WWW 画像検索システムにおいては，画像とリンク先のページ内容が必ず一致するとは限らない．したがって，画像のみを見てリンク先のページが適合文書か不適合文書であるかの判断することは困難である．例えば，図 1 では「松井」に関する画像を検索し，「松井秀喜が野球をしている画像」を正解画像としている．他の野球選手の画像も検索結果に含まれているが，これは不正解画像である．しかし，不正解画像となっている野球選手のページ中には，「松井」の類義語が多く含まれているため，このページを不適合文書とすると適合文書中に含まれる類義語の重要度を低下させてしまう．そこで，正解画像の言語情報のみを用いて以下の手順によって類義語を特定する．

手順 1: 類義語候補単語の重み付け
3.1 の手順 2 で得た類義語候補単語 w_j の周辺の HTML タグを利用して単語 w_j の重み付けを行う[3].

手順 2: 類義語候補単語が存在するページを検索
上位の w_j を WWW 画像検索システムに入力し，単語毎に上位 n 件の検索結果 URL を得る．

手順 3: 類義語候補単語の関連度を計算
検索結果 URL 群に対応する HTML を形態素解析し，単語を得る．この単語群に類義語候補単語がどれだけ含まれているかを調べ，式(1)により類義語を特定する．

関連度 = *URL*群に存在する他の類義語数 ×
*n*件の*URL*群に他の類義語が存在する*URL*数 (1)

図 2 に上記の手順に従い，類義語の特定を行った例を示す．いま，5 つの類義語候補単語があるとする．まず，類義語候補単語「松井秀喜」を WWW 画像検索システムに入力し，上位 n 件の検索結果 URL に出現する単語を取得する．このとき「ヤンキース」，「本塁打」，「大リーグ」の単語が得られたとする．次に，この単語中に他の 4 つの類義語候補単語が含まれているかを調べると「ヤンキース」と「本塁打」が含まれていることがわかる．最後に，n 件の URL に出現する類義語候補単語の総数と出現 URL 数から式(1)により関連度を求める．「本塁打」等の類義語候補単語は n 件の URL 中に 20 回出現し，18 件の URL に含まれていたとすると「松井秀喜」の検索単語との関連度は 360 となる．また，「記事」を検索質問として検索されるページには，他の類義語候補語がほとんど含まれていないため，関連度は低くなっている．このように本手法を適用すると「記事」のような一般的な単語を類義語候補から除去することができる．

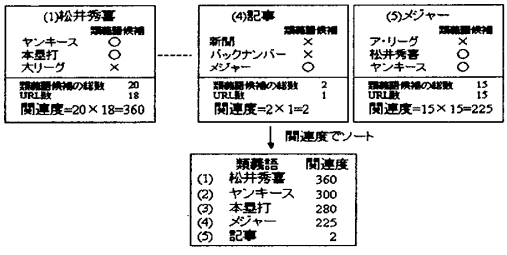

図 2： 類義語特定手順の例

4. 評価

4.1 評価条件

本稿で提案した類義語特定手法の有効性を確かめるために検索単語の類義語を収集して評価を行った．検索単語「松井」の検索結果画像 20 件から 7 件の画像を選択し，得られた 440 単語のうち HTML タグにより重み付けした単語の上位 100 単語を類義語候補単語として，本手法と適合性フィードバックにより類義語の特定を行った．また，類義語であるか否かの判断は人手により行い，精度評価には平均適合率を用いた．

4.2 結果

各手法により特定した類義語のうち上位から 25，50，75，100 位までの類義語を対象にして平均適合率を求めた．実験結果を図 3 に示す．

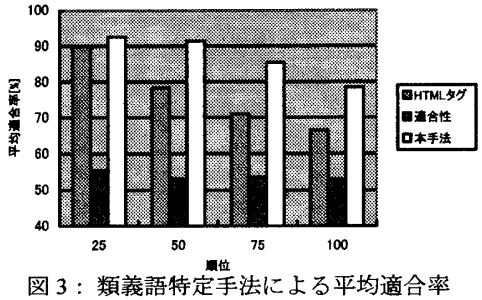

図 3： 類義語特定手法による平均適合率

実験結果より，本手法は，HTML タグで重み付けした結果よりもよい結果を得ていることがわかる．一方，適合性フィードバックを用いると上位 25 単語でも多くのノイズ単語を含んでおり，平均適合率 50%で推移している．これは，不正解画像としたページ内に多くの適切な類義語候補語を含んでいたためである．画像のみからページの内容を判断することは非常に困難であり，単純に不正解画像のページを不適合文書と見なすことはできず，適合性フィードバックを適用できないことが確認できた．

今回の実験では，類義語の特定精度についてのみ評価を行ったが，本手法で特定した上位の類義語を用いて再検索することで，正解画像に類似した画像がより多く収集できることが考えられる．

5. まとめ

本稿では，WWW 画像検索において，フィードバック情報を用いて検索質問の類義語を特定し，類義語を用いてフィードバック検索を行う手法を提案した．また，実データに対する実験を行い，本類義語特定手法の有効性を確認できた．今後は，特定した類義語を用いた画像検索実験を行い，本手法の有効性を更に検討したい．

謝辞：本研究の一部は，科学研究費補助金基盤研究(B)(17300036)，科学研究費補助金基盤研究(C)(17500644)を受けて行われた．

参考文献

[1]獅々堀正幹，小泉大地，柘植覚，北研二，“画像知識データベースを用いた WWW 画像検索システムの開発”，電子情報通信学会論文誌，VOL.J87-D-I NO.2，pp.154-163，2004.
[2]Rocchio，J. J.，“Relevance feedback in information retrieval”，The SMART Retrieval System-Experiments in Automatic Document Processing，Salton，G.(Ed)，Prentice Hall，pp.313-323，1971.
[3]杉尾敏康，竹野浩，藤本典幸，萩原兼一，“WWW に対するマルチメディアデータ検索エンジンの HTML 構文を活かしたスコア付け手法の提案”，第 13 回データ工学ワークショップ(DEWS2002)，2002.